MINIMAL NUISANCE, MAXIMAL FIDELITY: LATENT-STATE-CONDITIONED INVARIANCE IN DOMAIN GENERALISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalisation across physical systems remains difficult because domains differ in equipment, operating regimes, and system dynamics. As a result, models often rely on system-specific shortcuts rather than learning transferable process behaviour. Existing invariant representation learning methods typically enforce global domain invariance, or condition only on coarse task labels, potentially erasing useful within-label behavioural structure. We argue that the appropriate target is *latent-state-conditioned nuisance invariance* (LaSNI), where domain-specific variation should be removed only after conditioning on the underlying behavioural state. We formalise this target with $z \perp d \mid s$, where z is the learned representation, d the domain, and s an unobserved latent behavioural state. Since s is unobserved, we propose *Local Behavioural Nuisance Suppression* (LBNS), a source-only surrogate that constructs dynamic same-label neighbourhoods, estimates a local behavioural reference, and suppresses cross-domain variation in the residual rather than the full representation. Across six real-world tasks and 30 leave-domain-out folds, LBNS achieves the strongest aggregate transfer and lowest average domain recoverability across all eight objectives. An independent descriptor diagnostic shows that its latent neighbours are more behaviourally coherent than random same-label samples in every fold and dataset, including under cross-domain restriction. These results support conditioning nuisance suppression more finely than the observed class when labels hide heterogeneous behaviour.

1 INTRODUCTION

Physical systems are subject to substantial distribution shift across sites, assets, and operating environments. Even when the downstream task is nominally shared, e.g., condition monitoring, fault detection, or process-state classification, the observed measurements may differ significantly due to sensor heterogeneity, machine-specific response characteristics, calibration differences, maintenance history, and regime-dependent dynamics. These factors induce a domain shift that often causes models trained on one subset of systems to rely on spurious system-specific cues rather than transferable process behaviour. As a result, cross-system generalisation remains a challenge in industry, particularly when labels are coarse, sparse, or inconsistently expressed across assets (Chen et al., 2024; Zhao et al., 2024).

A common strategy for domain generalisation is to learn representations that are invariant across domains (Wang et al., 2023), either globally (Ganin et al., 2016; Jia & Zhang, 2024) or by conditioning alignment on the observed task label (Long et al., 2018; Nguyen et al., 2025). While such objectives can be effective in homogeneous settings, they are often too coarse for real-world industrial data. In practice, coarse labels such as *normal* and *fault* states can themselves aggregate multiple operating regimes, process phases, fault expressions, or machine-dependent manifestations of the same underlying state. Enforcing full domain invariance, or even coarse class-conditional invariance, can therefore over-align samples that should remain distinct, collapsing meaningful within-label structure by conflating it with nuisance variation. In this setting, the goal is not to remove all domain-dependent variation, but to suppress only the nuisance component while preserving task-relevant behavioural heterogeneity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We argue that the appropriate target is *latent-state-conditioned nuisance invariance* (LaSNI). Let z denote the learned representation, d the system or domain identifier, and s an unobserved latent behavioural state. Rather than seeking $z \perp d$ or $z \perp d \mid y$ for a coarse label y, we instead aim for

$$z \perp d \mid s.$$

This criterion states that once the underlying behavioural state is fixed, the representation should no longer reveal which system generated the sample. Intuitively, the model should discard nuisance domain information only within matched behavioural states, not globally across all samples sharing the same coarse label. This perspective preserves heterogeneous task structure while discouraging spurious domain shortcuts, making it especially well-suited to heterogeneous physical systems.

Because the latent state s is unobserved, we study a practical surrogate, *Local Behavioural Nuisance Suppression* (LBNS), which reduces cross-domain variation only among locally comparable same-label samples in latent space. The broader claim of this paper, however, is not that a particular local regulariser is universally optimal. Rather, we argue that the granularity at which nuisance suppression is conditioned matters, as global invariance can be too coarse, and even coarse class-conditional control may remain too coarse when observed labels hide substantial latent behavioural heterogeneity. Accordingly, we use LBNS as one tractable surrogate to evaluate our hypothesis and address the specific question *at what granularity nuisance suppression should be conditioned, and to which component of the representation it should be applied*.

Our contributions are as follows:

- We identify and formalise the conditioning-granularity failure mode: when the observed label $y = h(s)$ is a many-to-one function of the latent state, enforcing $z \perp d \mid y$ can suppress useful within-label structure; we formulate $z \perp d \mid s$ for an unobserved, potentially continuous state that need not be independent of the domain.
- We operationalise this target with LBNS, a source-only surrogate that separates each representation into a dynamic local behavioural reference and residual variation, suppressing cross-domain residual nuisance while preserving local within-class structure without target observations, state annotations, discretisation, or a predefined subtype count.
- We test the hypothesis through a controlled progression from no suppression to global, class-conditional, conditional-adversarial, and local suppression across six real-world time series tasks under realistic leave-domain-out evaluation, measuring transfer, recoverable domain information, and neighbourhood coherence.

2 Related Work

A central challenge in domain generalisation is heterogeneous real-world operating conditions (Jiang et al., 2025). Out-of-distribution generalisation in time series has recently emerged as an active research area, with prior work organised around data distribution shift, representation learning, and evaluation protocols (Wu et al., 2026). More broadly, domain generalisation methods commonly pursue robustness through invariant representation learning, discrepancy minimisation, adversarial alignment, and related feature-disentanglement strategies (Wang et al., 2023). Within this landscape, most approaches suppress domain information either globally or at the level of coarse task labels. A separate concern is whether standard, often well-balanced, benchmarks adequately capture the robustness challenges that arise in real deployments. Recent work has suggested that many popular domain generalisation benchmarks may not fully capture robustness to spurious correlations, as out-of-distribution shifts do not always sufficiently penalise reliance on shortcuts (Salaudeen et al., 2025). This is particularly relevant for industry, where shifts induced by operating conditions, sensing configurations, hardware variation, maintenance cycles, and evolving control policies are operational rather than purely benchmark-defined (Kamm et al., 2023; Eldele et al., 2024). Complementary approaches enforce invariance at the predictor, risk, or inferred-subdomain level. IRM seeks representations for which the same predictor is optimal across source environments (Arjovsky et al., 2019), V-REx penalises variation in environment risks (Krueger et al., 2021), and DIVERSIFY infers latent time series subdomains and adversarially suppresses their distribution gaps (Lu et al., 2023). In contrast, the latent variable in LaSNI denotes a task-relevant behavioural state to preserve, whereas LBNS conditions suppression on local behavioural context and targets only residual observed-domain variation.

Prior work has shown that global domain matching can be insufficient when the relevant conditional structure remains misaligned. Nguyen et al. (2021), for instance, show that enforcing marginal invariance alone may not resolve conditional mismatch, while Stojanov et al. (2021) argue that a single fixed encoder shared across domains can be suboptimal when domain-specific effects alter the input-label relationship. Other work addresses domain generalisation in temporally evolving settings (Cai et al., 2024) but focuses on modelling domain dynamics rather than the conditioning granularity of nuisance suppression. More broadly, Nastl & Hardt (2024) question whether stronger invariance assumptions necessarily yield better generalisation to new domains, suggesting that the effectiveness of invariant predictors depends on how the underlying shift is structured. MatchDG (Mahajan et al., 2021) showed that class-level conditioning can be insufficient when classes contain heterogeneous underlying objects. This reinforces our view that, in heterogeneous settings, the key issue is not whether nuisance variation should be suppressed, but at what granularity that suppression should occur and how. We address the continuous-state setting where exact object matching is unavailable and potentially undesirable, as locally comparable behavioural states can legitimately differ.

Fine-grained and neighbourhood-based alignment provide important antecedents. Subtype-aware dynamic UDA discovers subtype structure using unlabeled target observations and intermediate pseudo-labels (Liu et al., 2024); DSAN aligns source and target subdomain distributions with local maximum mean discrepancy and target-derived assignment weights (Zhu et al., 2021); and graph-embedding analyses of supervised domain adaptation connect source and target samples through pairwise neighbourhood relations (Hedegaard et al., 2021). These methods establish the value of structure below the class level, but they operate in adaptation settings where target observations participate in representation learning. Our setting is source-only domain generalisation where held-out domains are absent from representation learning, neighbourhood construction, hyperparameter selection, and checkpoint selection. LBNS additionally aligns deviations from a dynamic per-anchor local reference rather than complete representations, subtype prototypes, or target-aware subdomain distributions. Our claim is therefore not component-level novelty for neighbourhoods or conditional alignment; it is the formulation and empirical test of conditioning granularity as a source-only nuisance-suppression problem, together with a locally residualised surrogate.

These observations point to a specific gap in the current domain-generalisation and time series out-of-distribution literature: the conditioning granularity of nuisance suppression when the relevant state is unobserved and potentially continuous. We study whether source-only nuisance control should be conditioned more finely than the observed class label, rather than claiming that fine-grained alignment itself is unprecedented.

3 METHOD

We consider supervised classification under multiple source domains. Each example consists of an input time series $x_i \in \mathcal{X}$, a task label $y_i \in \{1, \dots, C\}$, and a domain label $d_i \in \{1, \dots, D\}$. A feature extractor g_θ maps the input to a latent representation (z_i) and a task head h_ϕ predicts the task label,

$$z_i = g_\theta(x_i) \in \mathbb{R}^H, \qquad \hat{y}_i = h_\phi(z_i).$$

Because the latent behavioural state is unobserved, we operationalise finer-grained nuisance suppression through local same-label neighbourhoods in latent space as illustrated in Figure 1. LBNS reduces cross-domain residual variation only among behaviourally comparable samples while preserving local within-class structure. These are combined with the standard supervised task loss, which gives

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}} + \lambda_{\text{struct}}\mathcal{L}_{\text{struct}},$$

where $\mathcal{L}_{\text{task}}$ is the supervised classification loss, $\mathcal{L}_{\text{inv}}$ is the local nuisance-suppression term, and $\mathcal{L}_{\text{struct}}$ is the local structure regulariser. Their respective strengths are controlled independently by λ_{inv} and λ_{struct}. For each sample i, we define a same-label local neighbourhood $\mathcal{N}_k(i)$ as the set of its k Nearest Neighbours (KNN) in latent space among examples with the same coarse task label, such that

$$\mathcal{N}_k(i) = \text{KNN}_k\big(z_i; \{z_j : y_j = y_i,\ j \neq i\}\big).$$

Therefore, neighbourhoods are restricted to task-compatible samples while still allowing multiple local modes within each coarse class. We assign similarity weights within each neighbourhood using

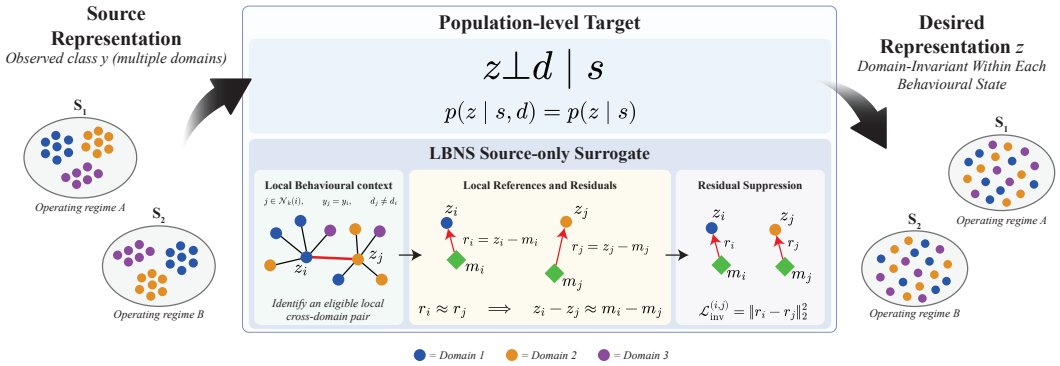


Figure 1: LaSNI preserves differences between latent behavioural states while suppressing within-state domain variation. Since s is unobserved, LBNS constructs same-label local references from source samples and aligns cross-domain residuals $r_i = z_i - m_i$. This yields $z_i - z_j \approx m_i - m_j$, preserving locally explained behavioural differences while reducing domain mismatch.

a Radial Basis Function (RBF) kernel, which defines a local behavioural reference as follows

$$w_{ij} = \frac{\exp\left(-\frac{\|z_i - z_j\|_2^2}{2\sigma^2}\right) \mathbf{1}[j \in \mathcal{N}_k(i)]}{\sum_{m=1}^{B} \exp\left(-\frac{\|z_i - z_m\|_2^2}{2\sigma^2}\right) \mathbf{1}[m \in \mathcal{N}_k(i)] + \varepsilon}, \qquad m_i = \sum_{j \in \mathcal{N}_k(i)} w_{ij} z_j, \qquad r_i = z_i - m_i,$$

where m_i summarises the shared local behaviour around sample i and r_i denotes the corresponding residual. Intuitively, m_i captures the shared local behaviour of samples similar to i, while r_i captures the remaining variation around that local behaviour. We treat cross-domain differences in these residuals as candidate nuisance variation. To suppress such nuisance structure only where comparison is meaningful, we penalise residual mismatch between samples that are (i) locally neighbouring, (ii) in the same coarse class, and (iii) from different domains by defining the comparison-pair set $\mathcal{P}$ as

$$\mathcal{P} = \{(i,j) : j \in \mathcal{N}_k(i),\ y_i = y_j,\ d_i \neq d_j\}, \qquad \mathcal{L}_{\text{inv}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \|r_i - r_j\|_2^2.$$

This objective differs from global-domain matching. Directly penalising $\|z_i - z_j\|_2^2$ conflates differences in local behavioural structure with residual variation and may therefore cause over-alignment. LBNS instead penalises $\|r_i - r_j\|_2^2 = \|(z_i - z_j) - (m_i - m_j)\|_2^2$, only for locally comparable, same-label, cross-domain pairs, while allowing $m_i \neq m_j$. Importantly, the domain-suppression term does not align complete representations; it aligns each sample's displacement from its own local behavioural reference. Consequently, suppressing residual mismatch preserves the difference explained by local behavioural structure while discouraging domain-specific variation beyond that structure. Applying only $\mathcal{L}_{\text{inv}}$ could lead to excessive smoothing or neighbourhood collapse. We therefore include a local structure term that encourages compactness among same-label neighbours:

$$\mathcal{L}_{\text{struct}} = \frac{1}{|\mathcal{Q}|} \sum_{(i,j) \in \mathcal{Q}} \|z_i - z_j\|_2^2, \qquad \mathcal{Q} = \{(i,j) : j \in \mathcal{N}_k(i),\ y_i = y_j\}.$$

This term encourages local compactness among task-compatible neighbours and stabilises the representation against arbitrary distortions. Therefore, the total loss is

$$\mathcal{L} = \underbrace{\frac{1}{B} \sum_{i=1}^{B} \ell_{\text{CE}}(h_\phi(z_i), y_i)}_{\mathcal{L}_{\text{task}}} + \lambda_{\text{inv}} \underbrace{\frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \|r_i - r_j\|_2^2}_{\mathcal{L}_{\text{inv}}} + \lambda_{\text{struct}} \underbrace{\frac{1}{|\mathcal{Q}|} \sum_{(i,j) \in \mathcal{Q}} \|z_i - z_j\|_2^2}_{\mathcal{L}_{\text{struct}}}.$$

To make the training procedure explicit, Algorithm 1 in Appendix A.2 describes one LBNS update. We recompute neighbourhoods from the current representation in every minibatch rather than storing them in a fixed graph. Our implementation forms the dense $B \times B$ distance matrix and therefore has loss-side complexity $\mathcal{O}(B^2 H)$ for minibatch size B and representation dimension H; local-reference and loss computation after neighbour selection costs $\mathcal{O}(BkH)$. LBNS adds no trainable parameters, and its cost depends on the minibatch rather than the full dataset.

4 THEORETICAL MOTIVATION

We now formalise the representation-learning target that motivates our method. The results in this section are intended as theoretical motivation rather than as general target-domain risk guarantees. We provide proof sketches in the main text for readability, with full formal statements and proofs in Appendix B.

Definition 1 (Global domain invariance)**.** *A representation z satisfies global domain invariance if $p(z \mid d)$ is invariant with respect to d.*

Definition 2 (Class-conditional domain invariance)**.** *A representation z satisfies class-conditional domain invariance if $p(z \mid y, d)$ is invariant with respect to d.*

Definition 3 (Latent-state-conditioned nuisance invariance)**.** *Let s denote an unobserved latent behavioural state. A representation z satisfies LaSNI if*

$$z \perp d \mid s \quad \iff \quad p(z \mid s, d) = p(z \mid s).$$

This definition should be understood as the population-level target, not as a property we can verify directly. Since the latent behavioural state is unobserved, we approximate conditioning on s through local neighbourhoods of behaviourally similar samples in representation space.

LaSNI does not assume $s \perp d$. The conditional statement permits arbitrary statistical or causal dependence between behavioural state and domain. If a machine or maintenance history affects wear state, that realised state belongs to s and should be preserved; LaSNI suppresses only domain variation that remains after the state is fixed. The substantive premise is therefore the existence of some removable residual nuisance within comparable states, not domain-independence of the states themselves.

Assumption 1. *There exists a latent behavioural state s such that the observed coarse label y is a deterministic many-to-one function of s, that is, $y = h(s)$ for some mapping $h : \mathcal{S} \to \mathcal{Y}$.*

Assumption 2. *Distinct latent states sharing the same coarse label may exhibit different task-relevant dynamics, so that preserving information about s is necessary for prediction.*

Assumption 3. *For each latent behavioural state s, domain variation contains a nuisance component that is, in principle, removable after conditioning on s.*

Assumption 4 (Local behavioural coherence)**.** *Within each observed class, same-label neighbourhoods in representation space are more homogeneous in latent behavioural state than the class as a whole. Thus, if $j \in \mathcal{N}_k(i)$ and $y_j = y_i$, then s_j is a finer approximation to s_i than an arbitrary sample drawn from class y_i.*

Proposition 1. *Under Assumptions 1–3, class-conditional invariance with respect to the coarse label y can be too coarse: enforcing $z \perp d \mid y$ may suppress domain-dependent variation that is also informative about distinctions among latent states s within the same label. By contrast, $z \perp d \mid s$ removes only residual domain-specific nuisance variation after conditioning on the behavioural state and can therefore preserve within-label task-relevant structure.*

Proof sketch. Consider two latent states within the same observed class that occur in both domains but in different proportions. For example, let $\Pr(u = 1 \mid d = 0, y = 0) = 0.25$ and $\Pr(u = 1 \mid d = 1, y = 0) = 0.75$, where u distinguishes the two states. If a representation preserves this distinction, its class-conditional distribution differs across domains because the state mixtures differ. Enforcing $z \perp d \mid y$ therefore requires removing that within-label state distinction in this construction. By contrast, the representation $z = s$ satisfies $z \perp d \mid s$ while preserving the state, and this statement is non-vacuous because both states occur in both domains. A formal construction and proof are provided in Appendix B.1. □

Proposition 2. *Under Assumptions 1–4 and the local regularity conditions stated in Appendix B.2, the LBNS residual mismatch for a cross-domain local pair approximates the within-state cross-domain nuisance discrepancy up to bounded neighbourhood, reference-estimation, and noise errors. By contrast, direct representation mismatch also contains the behavioural displacement between locally comparable but potentially non-identical states. Thus, the residual-suppression term provides a tractable surrogate for locally conditioned nuisance suppression without requiring exact latent-state matches.*

Proof sketch. Because s is unobserved, $z \perp d \mid s$ cannot be enforced directly. Assumption 4 motivates using local same-label pairs as approximate rather than exact state matches. Under the local decomposition $z_i = a(s_i) + n(s_i, d_i) + \epsilon_i$ and a bounded local-reference error $m_i = a(s_i) + e_i$, the residual satisfies $r_i = n(s_i, d_i) + \epsilon_i - e_i$. For nearby states, local smoothness of n then bounds the difference between $r_i - r_j$ and the within-state nuisance discrepancy by a term that depends on state mismatch, reference error, and noise. Direct representation mismatch additionally contains $a(s_i) - a(s_j)$, which need not vanish for nearby but non-identical states. The formal bound and proof are provided in Appendix B.2. □

5 EXPERIMENTS

To evaluate our method, we use six real-world operational time series tasks spanning industrial fault detection, electricity demand, air quality, and traffic monitoring:

PREDIST A district heating dataset with 93 substation domains at 10-minute resolution and two labelled subsets, `M1` and `M2` (Roelofs et al., 2025). We derive fault labels from incident reports and maintenance data. `M1` contains 291 daily samples, including 56 fault days, and `M2` contains 404 daily samples, including 124 fault days.

EC A fault detection field dataset from 36 energy centres at 5-minute resolution, each corresponding to a renovated collective boiler-room installation for apartment buildings (van Dreven et al., 2026). It contains 2040 daily samples, including 487 fault days.

ELD The Electricity Load Diagrams dataset, comprising 370 client domains with 15-minute electricity load measurements in kW (Trindade, 2015). We predict a binary high-variability vs low-variability task based on within-day load-profile variability. We use 3789 daily samples in total, of which 1905 belong to the low-variability class.

BMSAQ The Beijing Multi-Site Air Quality dataset, comprising 12 monitoring-site domains with hourly measurements (Chen, 2017). We use a high-vs-low PM2.5 classification task and restrict inputs to O3 to avoid trivial same-day recovery from highly correlated covariates, yielding a more meaningful cross-domain task. We use 3641 daily samples in total, of which 2424 belong to the high-PM2.5 class.

METR-LA A traffic time series dataset with 207 sensor domains and 3726 daily samples at 5-minute resolution from the Los Angeles traffic monitoring system (Li et al., 2018). Labels include morning vs. evening traffic, yielding a balanced split of 1863 samples per class.

We compare against structured comparators that span progressively finer levels of nuisance control. Rather than benchmarking the full space of time series domain-generalisation methods, we aim to isolate the invariance objective. For all methods, we use the same Time Series Transformer (TST) (Zerveas et al., 2021), specified in Appendix A.3. We compare LBNS against ERM, DANN, CCA, CDANN, MatchDG, the time series OOD method DIVERSIFY (Lu et al., 2023), and the risk-invariance method V-REx (Krueger et al., 2021). All methods use the same encoder architecture and source-validation model-selection protocol; Appendix A.1 provides exact objectives and implementation details. For every dataset, each fold holds out previously unseen domains for final evaluation; training, hyperparameter selection, and checkpoint selection use only the remaining source domains. We draw the validation split exclusively from those sources and select checkpoints by source-validation ROC AUC. Models are trained from scratch with a batch size of 32 and Adam at 10^{-3} without weight decay. We train M1, M2, ELD, and BMSAQ for 30 epochs and EC and METR-LA for 8 epochs. We report ROC AUC and PR AUC over 30 paired folds. Omnibus comparisons use Friedman tests; post-hoc comparisons use paired Wilcoxon signed-rank tests with Holm correction, and paired effect sizes use Cohen's d_z. Appendix A includes full objectives, hyperparameters, sensitivity analyses, and probe details.

6 RESULTS

Figure 2 and Table 1 report target-domain performance. The eight-method Friedman test is significant for all 12 dataset–metric comparisons ($p \leq 0.015$). LBNS obtains the highest mean ROC and PR AUC on M1, M2, EC, and METR-LA; V-REx is narrowly highest on ELD and MatchDG on

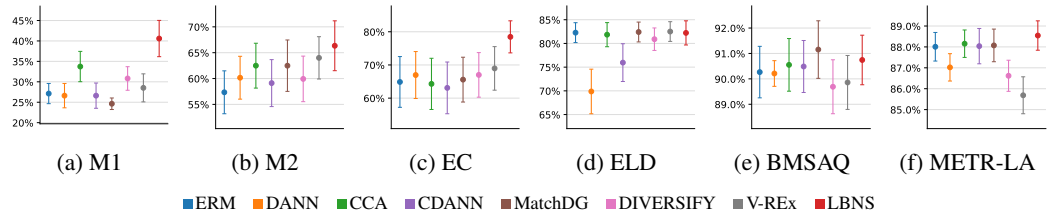


Figure 2: Target-domain PR AUC under 30-fold leave-domain-out evaluation. Points show fold means and error bars show 95% confidence intervals. The panels use dataset-specific y-axis scale. All objectives use the same encoder and source-only model-selection protocol. ROC AUC is reported in Table 1.

Method	M1	M2	EC	ELD	BMSAQ	METR-LA	Avg.
ERM	44.1±3.0	51.8±5.1	74.5±5.0	82.0±2.2	83.9±0.8	88.6±0.6	70.8
DANN	45.1±3.2	55.5±5.1	75.7±5.0	69.0±4.6	84.1±0.6	87.7±0.6	69.5
CCA	50.6±2.6	60.7±5.0	74.0±5.3	82.1±2.4	84.0±0.8	88.7±0.5	73.4
CDANN	43.8±3.2	53.6±5.5	73.7±5.0	74.8±3.8	84.1±0.8	88.7±0.7	69.8
MatchDG	41.4±2.5	58.1±5.6	75.7±4.6	82.4±1.9	**85.3±0.9**	88.8±0.5	72.0
DIVERSIFY	51.6±3.2	54.7±5.8	76.7±4.4	80.7±2.5	83.2±0.8	87.0±0.7	72.3
V-REx	46.4±4.5	60.9±3.3	76.9±4.5	**82.5±2.1**	83.0±0.7	85.8±0.8	72.6
LBNS	**57.5±3.6**	**64.7±4.9**	**81.1±3.9**	82.1±2.4	84.6±0.8	**89.0±0.5**	**76.5**

Table 1: Target-domain ROC AUC (%) with 95% confidence intervals over 30 folds. The final column is the unweighted dataset average.

BMSAQ. LBNS exceeds DIVERSIFY on both metrics on all six tasks and V-REx on five of six; 9/12 paired gains against each method remain significant after Holm correction. LBNS also has the best average fold rank across all dataset–metric comparisons (3.06; next-best MatchDG 3.90 and CCA 4.01). This pattern supports the conditioning-granularity hypothesis: latent-subdomain discovery, risk invariance, label-conditioned adversarial suppression, and complete-output matching do not recover the aggregate gains of local behavioural conditioning. The effect size varies with physical heterogeneity and available headroom. M1, M2, and EC contain distinct district heating substations or energy centres whose equipment, control, sensing, occupancy, usage, and maintenance differ; these tasks show the largest improvements. ELD, BMSAQ, and METR-LA use more standardised sensing equipment within a common deployment context, leaving less heterogeneous nuisance to remove. Accordingly, several methods are effectively tied on ELD; ERM already has relatively low domain recoverability on BMSAQ and reaches 88.6 ROC AUC on METR-LA. LBNS remains competitive-to-best in these milder settings without sacrificing predictive performance.

As shown in Figure 3 and Table 2, LBNS achieves the lowest average domain recoverability (2.071 times chance; next-best CCA 3.478). Its ratio is lower than CDANN, MatchDG, DIVERSIFY, and V-REx on all six tasks; against DIVERSIFY and V-REx, every fold-paired difference remains significant after Holm correction (adjusted $p < 0.001$). Thus, neither observed-label conditioning, matched-output regularisation, latent-subdomain discovery, nor risk invariance explains the result. This supports the central hypothesis in two ways. First, reducing recoverable domain information is generally associated with improved transfer to unseen domains. Second, the strongest results come not from broad global or class-conditional adversarial suppression, but from finer-grained nuisance control that removes domain-specific variation only among behaviourally comparable samples, preserving within-class structure useful for prediction. Moreover, we observe that stronger adversarial pressure for DANN reduces the accuracy of the co-trained domain head during optimisation, suggesting increasing domain confusion. However, post-hoc probes still recover substantial domain information. This reveals a failure mode in which the encoder can fool the specific adversary used during training without removing system-specific structure from the final representation. Our observations align with prior findings that adversarial removal of an attribute does not necessarily imply the attribute is absent from the learned representation (Elazar & Goldberg, 2018; Kumar et al., 2022), while extending this concern to time series domain generalisation. We verified that this behaviour persists

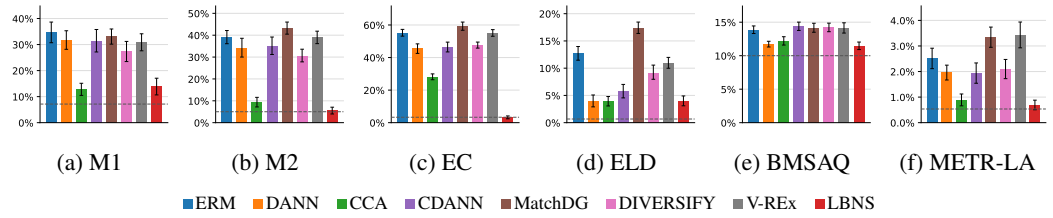


Figure 3: Post-hoc source-domain accuracy on held-out source-validation representations. Each panel uses the dataset's raw accuracy scale. The dashed line marks random-chance accuracy $1/|\mathcal{D}_{\text{source}}|$. Higher indicates more recoverable domain information; accuracy near the chance line indicates lower recoverable source-domain information. Error bars are 95% confidence intervals over 30 folds.

Method	M1	M2	EC	ELD	BMSAQ	METR-LA	Avg. Ratio
ERM	4.853	7.825	16.567	18.957	1.388	4.698	9.048
DANN	4.443	6.860	13.644	5.946	1.172	3.664	5.955
CCA	**1.792**	1.877	8.450	**5.862**	1.221	1.665	3.478
CDANN	4.405	7.035	13.939	8.618	1.439	3.625	6.510
MatchDG	4.629	8.649	17.783	25.953	1.417	6.252	10.781
DIVERSIFY	3.827	6.113	14.300	13.655	1.423	3.903	7.203
V-REx	4.331	7.781	16.539	16.341	1.416	6.418	8.804
LBNS	1.941	**1.105**	**0.978**	5.960	**1.148**	**1.295**	**2.071**

Table 2: Held-out source-validation domain-probe accuracy normalised by source-domain random-chance accuracy ($1/|\mathcal{D}_{\text{source}}|$). Lower is better, with 1 indicating chance-level domain recoverability.

under substantially stronger adversarial settings, including much larger domain-loss and gradient-reversal coefficients, indicating that the effect cannot be attributed to a weakly tuned adversarial baseline. Therefore, low training-time adversarial accuracy can create the appearance of invariance without eliminating recoverable domain information in the learned representation, and the domain information may be too entangled with task-relevant industrial structure to remove globally without degrading prediction. The post-hoc probe measures global source-domain recoverability and is therefore complementary evidence rather than a direct test of $z \perp d \mid s$.

Next, we measure the effect of residual alignment vs direct representation alignment (full ablation in Appendix A.7). LBNS improves PR on all $6/6$ datasets with a large effect size (paired Wilcoxon $p = 0.031$; $d_z = 0.84$) and ROC on $5/6$ (paired Wilcoxon $p = 0.094$; $d_z = 0.89$); domain-probe accuracy was statistically indistinguishable ($p = 0.844$; $d_z = 0.133$). Thus, direct representation alignment removes a comparable amount of recoverable domain information but yields worse predictive performance, particularly on the heterogeneous datasets. Residual alignment therefore provides a better invariance–fidelity trade-off by preserving locally explained behavioural structure. In a matched ablation, we remove the structure term. Full LBNS improves PR AUC on 5/6 datasets by 1.46 points on average (paired Wilcoxon $p = 0.094$; $d_z = 0.76$) and ROC AUC on 4/6 datasets by 0.65 points (paired Wilcoxon $p = 0.219$; $d_z = 0.76$), with the largest benefit on EC ($+5.02$ PR points). Domain-probe accuracy remains statistically indistinguishable ($p = 0.563$; $d_z = 0.32$), suggesting that residual alignment supplies the central invariance mechanism, while the structure term acts as a stabilising fidelity regulariser rather than merely increasing domain suppression. Moreover, we retrained LBNS at $H \in \{16, 64\}$ over all 30 folds to test representation-dimension sensitivity. $H = 16$ changes mean ROC AUC, PR AUC, and source-validation probe accuracy by -1.80, -1.68, and $+1.67$ percentage points, respectively, while $H = 64$ changes them by -0.65, -1.05, and -0.46 points; none of the six comparisons remains significant after Holm correction (paired Wilcoxon, adjusted $p \geq 0.188$), suggesting a stable pattern across dimensions. We report complete results in Appendix A.6. To examine the local-conditioning premise more directly, we next evaluate whether the learned neighbourhoods are behaviourally coherent. Because s is unobserved, we cannot measure true latent-state purity directly. We instead test Assumption 4's observable implication using descriptors that are not used by LBNS for neighbour selection. For each source-training series at every final LBNS checkpoint, we compute signal level, variability, trend, temporal roughness,

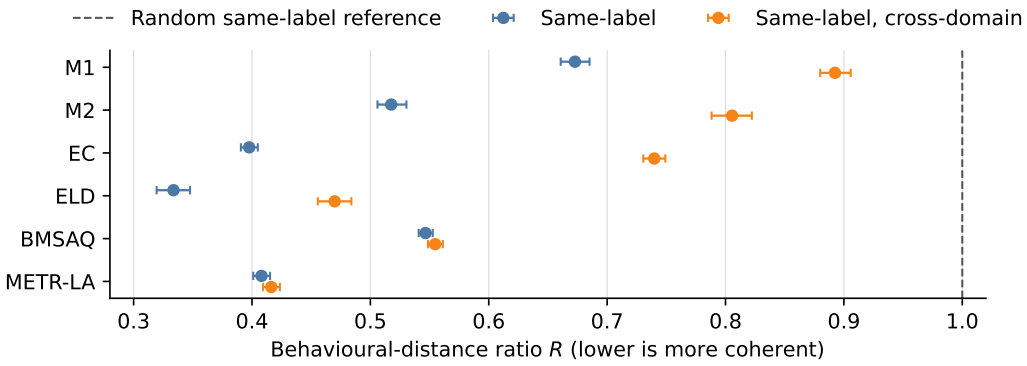


Figure 4: Behavioural coherence of LBNS neighbourhoods. Points show the ratio between descriptor distance to latent neighbours and the exact same-label random expectation; lower is more coherent and the dashed line is the random reference. Error bars are 95% fold-bootstrap intervals over 30 folds.

lag-one autocorrelation, and spectral-shape descriptors, robustly standardised within each fold. We compare the descriptor distance to the five nearest same-label latent neighbours against the exact expected distance under uniform same-label sampling. Figure 4 reports $R = D_{\mathrm{NN}}/D_{\mathrm{random}}$; $R < 1$ means the latent neighbours are more behaviourally coherent than arbitrary samples from the same class. We repeat the analysis after restricting both neighbours and controls to different domains. The ratio is below one in all 30 folds of every task, including under cross-domain restriction (one-sided fold-level sign-flip test $p < 0.001$ in each task), and the conclusion holds within each task label. Under the stricter cross-domain comparison, behavioural distance falls by 10.8% on M1, 19.4% on M2, 26.0% on EC, 53.0% on ELD, 44.5% on BMSAQ, and 58.4% on METR-LA relative to random same-label matching. On EC, LBNS is also more coherent than ERM, DANN, and CCA for both same-label and cross-domain neighbourhoods; all paired intervals exclude zero, with paired effect-size magnitudes from $|d_z| = 0.73$ to 1.50 and Holm-adjusted $p < 0.001$. This does not prove recovery of s, but it directly supports the operational assumption that local latent neighbourhoods provide a finer behavioural match than the observed class alone.

6.1 Limitations

LBNS is a practical surrogate for LaSNI rather than a proof that $z \perp d \mid s$. Because s is unobserved, the descriptor analysis supports neighbourhood coherence but cannot establish latent-state recovery, and representation-dependent neighbourhoods may remain imperfect. Finally, evaluation is limited to binary tasks and a single architecture family; multiclass tasks and broader architectures remain future work.

7 Conclusion

Standard domain-invariance objectives are often too coarse for heterogeneous settings, where coarse labels can aggregate multiple latent behavioural modes and domain-specific variation is partly entangled with task-relevant structure. This makes both global invariance and coarse class-conditional invariance prone to removing useful variation. We therefore formulate latent-state-conditioned nuisance invariance (LaSNI), $z \perp d \mid s$, as a finer-grained target and operationalise it through Local Behavioural Nuisance Suppression (LBNS), a source-only surrogate for finer-grained nuisance control that suppresses cross-domain variation only among behaviourally comparable samples defined through local within-class neighbourhoods in latent space. Across leave-domain-out experiments, the results show a consistent pattern, where global invariance objectives are often too coarse, structured suppression is clearly beneficial, and finer-grained control achieves the strongest overall transfer performance while also achieving the lowest average domain recoverability. Taken together, these findings support our hypothesis that, in heterogeneous settings, the key question in domain generalisation is not only whether to suppress domain information, but at what conditioning granularity and in which component of the representation.

AI USE STATEMENT

Generative AI tools were used solely to improve language and readability. The authors reviewed all edits and take responsibility for the final manuscript. No generative AI tools were used for other tasks requiring disclosure under the ICLR 2027 AI Policy.

REPRODUCIBILITY STATEMENT

Section 3 specifies the method and objectives, while the experimental section describes the leave-domain-out protocol, architecture, optimisation, baselines, and statistical analysis. The appendix reports the complete loss functions and coefficients, hyperparameters, probe protocol, sensitivity analyses, and component controls.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Zekun Cai, Guangji Bai, Renhe Jiang, Xuan Song, and Liang Zhao. Continuous temporal domain generalization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pp. 127987–128014, Red Hook, NY, USA, December 2024. Curran Associates Inc.

Hongtian Chen, Hao Luo, Biao Huang, Bin Jiang, and Okyay Kaynak. Transfer learning-motivated intelligent fault diagnosis designs: A survey, insights, and perspectives. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):2969–2983, 2024. doi: 10.1109/TNNLS.2023.3290974.

Song Chen. Beijing multi-site air quality. UCI Machine Learning Repository, 2017. doi: 10.24432/C5RK5G.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1002.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Label-efficient time series representation learning: A review. *IEEE Transactions on Artificial Intelligence*, 5(12):6027–6042, December 2024. doi: 10.1109/TAI.2024.3430236.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2030–2096, January 2016.

Lukas Hedegaard, Omar Ali Sheikh-Omar, and Alexandros Iosifidis. Supervised domain adaptation: A graph embedding perspective and a rectified experimental protocol. *IEEE Transactions on Image Processing*, 30:8619–8631, 2021. doi: 10.1109/TIP.2021.3118978.

Chen Jia and Yue Zhang. Meta-learning the invariant representation for domain generalization. *Machine Learning*, 113(4):1661–1681, April 2024. ISSN 1573-0565. doi: 10.1007/s10994-022-06256-y.

Yushan Jiang, Kanghui Ning, Zijie Pan, Xuyang Shen, Jingchao Ni, Wenchao Yu, Anderson Schneider, Haifeng Chen, Yuriy Nevmyvaka, and Dongjin Song. Multi-modal time series analysis: A tutorial and survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 6043–6053, New York, NY, USA, August 2025. Association for Computing Machinery. doi: 10.1145/3711896.3736567.

Simon Kamm, Sushma Sri Veekati, Timo Müller, Nasser Jazdi, and Michael Weyrich. A survey on machine learning based analysis of heterogeneous data in industrial automation. *Computers in Industry*, 149:103930, August 2023. doi: 10.1016/j.compind.2023.103930.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. Probing classifiers are unreliable for concept removal and detection. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 17994–18008, Red Hook, NY, USA, November 2022. Curran Associates Inc.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR '18)*, 2018.

Xiaofeng Liu, Fangxu Xing, Jane You, Jun Lu, C.-C. Jay Kuo, Georges El Fakhri, and Jonghye Woo. Subtype-aware dynamic unsupervised domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):2820–2834, 2024. doi: 10.1109/TNNLS.2022.3192315.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 1647–1657, Red Hook, NY, USA, December 2018. Curran Associates Inc.

Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. Out-of-distribution representation learning for time series classification. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=gUZWOE42l6Q.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7313–7324. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/mahajan21b.html.

Vivian Y. Nastl and Moritz Hardt. Do causal predictors generalize better to new domains? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pp. 31202–31315, Red Hook, NY, USA, December 2024. Curran Associates Inc.

A. Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5264–5275. Curran Associates, Inc., 2021.

Anh T. Nguyen, Lam Tran, Anh Tong, Tuan-Duy H. Nguyen, and Toan Tran. Casual: conditional support alignment for domain adaptation with label shift. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, volume 39, pp. 19668–19676. AAAI Press, 2025. doi: 10.1609/aaai.v39i18.34166.

C. Roelofs, E. Bastidas Guevara, T. Hugo, S. Faulstich, and A. Cadenbach. Predist dataset - operational data of district heating substations labelled with faults and maintenance information. Zenodo, November 2025. doi: 10.5281/zenodo.17522255.

Olawale Elijah Salaudeen, Nicole Chiou, Shiny Weng, and Sanmi Koyejo. Are domain generalization benchmarks with accuracy on the line misspecified? *Transactions on Machine Learning Research*, 2025.

Petar Stojanov, Zijian Li, Mingming Gong, Ruichu Cai, Jaime Carbonell, and Kun Zhang. Domain adaptation with invariant representation learning: What transformations to learn? In *Advances in Neural Information Processing Systems*, volume 34, pp. 24791–24803. Curran Associates, Inc., 2021.

Artur Trindade. Electricityloaddiagrams20112014. UCI Machine Learning Repository, 2015. doi: 10.24432/C58C86.

Jonne van Dreven, Abbas Cheddad, Ahmad Nauman Ghazi, Sadi Alawadi, Jad Al Koussa, Dirk Vanhoudt, Hoda Kamali, Jonas Cleiren, and Ivan Verhaert. From lab to practice: Towards a substation-invariant self-supervised transformer for low-label fault detection in district heating. *Energy and AI*, 25:100824, 2026. ISSN 2666-5468. doi: 10.1016/j.egyai.2026.100824.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, August 2023. doi: 10.1109/TKDE.2022.3178128.

Xin Wu, Fei Teng, Xingwang Li, Ji Zhang, Qiang Duan, and Tianrui Li. Out-of-distribution generalization in time series: A survey. *Information Fusion*, pp. 104336, April 2026. doi: 10.1016/j.inffus.2026.104336.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*, KDD '21, pp. 2114–2124, New York, NY, USA, August 2021. Association for Computing Machinery. doi: 10.1145/3447548.3467401.

Chao Zhao, Enrico Zio, and Weiming Shen. Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study. *Reliability Engineering System Safety*, 245:109964, 2024. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2024.109964. URL https://www.sciencedirect.com/science/article/pii/S0951832024000395.

Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1713–1722, 2021. doi: 10.1109/TNNLS.2020.2988928.

A Experimental and reproducibility details

A.1 Exact comparison objectives

All methods use the same encoder architecture and task head. Let $z_i = f_\theta(x_i)$ and let

$$\mathcal{L}_{\text{task}} = \frac{1}{B}\sum_{i=1}^{B} \text{CE}(p_\omega(y_i \mid z_i), y_i)$$

denote the supervised task objective.

ERM. ERM uses only the supervised objective without an explicit domain-invariance objective:

$$\mathcal{L}_{\text{ERM}} = \mathcal{L}_{\text{task}}.$$

DANN. DANN, adversarially suppresses domain information across the complete representation, and adds a domain head $H \to H \to H/2 \to D$ with ReLU hidden layers:

$$\mathcal{L}_{\text{DANN}} = \mathcal{L}_{\text{task}} + \frac{\lambda_D}{B \log D}\sum_{i=1}^{B} \text{CE}\left(q_\psi(\text{GRL}(z_i)), d_i\right),$$

where $\lambda_D = 0.8$. The gradient-reversal layer is the identity in the forward pass and multiplies the gradient received by the encoder by -1.

CDANN. Label-Conditioned DANN conditions adversarial domain prediction on the observed source label, and uses the same domain head but conditions its input on the observed source label:

$$\mathcal{L}_{\text{CDANN}} = \mathcal{L}_{\text{task}} + \frac{\lambda_C}{B \log D}\sum_{i=1}^{B} \text{CE}\left(q_\psi(\text{GRL}(z_i) + e_{y_i}), d_i\right),$$

where $e_{y_i} \in \mathbb{R}^H$ is a learned label embedding. This is a matched source-only conditional-adversarial control, not a target-aware protocol such as CDAN (Long et al., 2018).

CCA. CCA provides a coarse class-conditional alignment control. Let

$$a_i = \mathrm{GELU}(W_h z_i + b_h) \in \mathbb{R}^H$$

denote the hidden task-head representation immediately before the final binary classifier, and let $w_0, w_1 \in \mathbb{R}^H$ be the columns of its weight matrix. We define the classifier direction and the corresponding orthogonal representation as

$$u = \frac{w_1 - w_0}{\|w_1 - w_0\|_2}, \qquad a_i^{\perp} = a_i - (a_i^{\top} u)u.$$

For each nonempty class-domain cell $(c, d) \in \mathcal{G}$, let μ_{cd} be the mean of $a_i^{\perp}$ in that cell and let $\bar{\mu}_c$ be the mean of the available domain means for class c. The objective is

$$\mathcal{L}_{\mathrm{CCA}} = \mathcal{L}_{\mathrm{task}} + \frac{\alpha}{|\mathcal{G}|} \sum_{(c,d)\in\mathcal{G}} \|\mu_{cd} - \bar{\mu}_c\|_2^2.$$

Thus, CCA suppresses coarse domain-level variation within each observed class while preserving the classifier direction.

DIVERSIFY. DIVERSIFY (Lu et al., 2023) iterates three source-only stages: updating the encoder with joint class–pseudo-domain labels, refining latent-domain assignments by class-independent adversarial characterisation and cosine-centroid assignment, and learning a task representation adversarially invariant to the inferred domains. We retain its method-specific bottlenecks and auxiliary heads while replacing only the backbone with the shared TST encoder.

V-REx. V-REx (Krueger et al., 2021) penalises variation in source-domain risks. For the source-domain risks R_d, its objective is

$$\mathcal{L}_{\text{V-REx}} = \frac{1}{D} \sum_{d=1}^{D} R_d + \beta \,\mathrm{Var}(R_1, \ldots, R_D).$$

MatchDG. MatchDG first infers same-label cross-domain matches and then encourages matched samples to produce similar task outputs. In Phase I, each anchor i is assigned a same-label sample $\pi(i)$ from a different source domain. Matches are initialised randomly and updated every five epochs to the nearest eligible sample in the learned representation. With cosine similarity $\mathrm{sim}(\cdot,\cdot)$ and temperature $\tau = 0.05$, the contrastive objective is

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{|\mathcal{A}|} \sum_{i\in\mathcal{A}} \log \frac{\exp\left(\mathrm{sim}(z_i, z_{\pi(i)})/\tau\right)}{\exp\left(\mathrm{sim}(z_i, z_{\pi(i)})/\tau\right) + \sum_{n\in\mathcal{N}_i} \exp\left(\mathrm{sim}(z_i, z_n)/\tau\right)},$$

where $\mathcal{N}_i$ contains samples in the current pair batch whose observed label differs from y_i. Anchors are sampled with source-domain balancing, and Phase I is trained for 20 epochs.

In Phase II, a fresh model is trained from scratch using the fixed matches inferred in Phase I. Let $\ell_i = g_\omega(f_\theta(x_i))$ denote the task logits, let $\mathcal{B}$ be an ordinary training minibatch, and let $\mathcal{P}$ be a minibatch of inferred same-label cross-domain pairs. Following the MatchDG objective, supervised cross-entropy is applied both to the ordinary minibatch and to both endpoints of each matched pair:

$$\mathcal{L}_{\mathrm{task}}^{\mathrm{ordinary}} = \frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \mathrm{CE}(\ell_i, y_i),$$

$$\mathcal{L}_{\mathrm{task}}^{\mathrm{matched}} = \frac{1}{2|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left[\mathrm{CE}(\ell_i, y_i) + \mathrm{CE}(\ell_j, y_j)\right].$$

The complete Phase-II objective is

$$\mathcal{L}_{\mathrm{MatchDG}} = \mathcal{L}_{\mathrm{task}}^{\mathrm{ordinary}} + \mathcal{L}_{\mathrm{task}}^{\mathrm{matched}} + \frac{\lambda_M(t)}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \|\ell_i - \ell_j\|_2^2,$$

where the coefficient is linearly increased from zero to λ_M over training. We select $\lambda_M \in \{0.01, 0.1, 1.0\}$ using source-validation performance only; neither matching nor model selection uses held-out target observations.

Algorithm 1 Local Behavioural Nuisance Suppression (LBNS)

Require: Minibatch $\mathcal{B} = \{(x_i, y_i, d_i)\}_{i=1}^{B}$, encoder g_θ, classifier h_ϕ, hyperparameters $k, \sigma, \lambda_{\text{inv}}, \lambda_{\text{struct}}$
Compute representations $z_i = g_\theta(x_i)$ and predictions $\hat{y}_i = h_\phi(z_i)$
Compute supervised loss $\mathcal{L}_{task} = \frac{1}{B}\sum_i \ell_{\text{CE}}(\hat{y}_i, y_i)$
for each $i \in \mathcal{B}$ **do**
    Construct same-label neighbourhood $\mathcal{N}_k(i)$ using k-NN in latent space
    Compute RBF weights w_{ij} for $j \in \mathcal{N}_k(i)$
    Compute local reference $m_i = \sum_{j\in\mathcal{N}_k(i)} w_{ij} z_j$ and residual $r_i = z_i - m_i$
end for
Form local pairs $\mathcal{Q} = \{(i,j) : j \in \mathcal{N}_k(i), y_i = y_j\}$
Form cross-domain pairs $\mathcal{P} = \{(i,j) \in \mathcal{Q} : d_i \neq d_j\}$
Compute $\mathcal{L}_{inv} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \|r_i - r_j\|_2^2$
Compute $\mathcal{L}_{struct} = \frac{1}{|\mathcal{Q}|}\sum_{(i,j)\in\mathcal{Q}} \|z_i - z_j\|_2^2$
Update θ, ϕ using $\mathcal{L} = \mathcal{L}_{task} + \lambda_{\text{inv}}\mathcal{L}_{inv} + \lambda_{\text{struct}}\mathcal{L}_{struct}$

A.2 Algorithm

A.3 Shared architecture and training protocol

All methods use the same Time Series Transformer encoder with three Transformer blocks, two attention heads, representation dimension $H = 32$, feed-forward dimension 64, dropout 0.3, learned positional embeddings, final layer normalisation, and global-average temporal pooling. The shared task head is a $32 \to 32 \to 2$ MLP with GELU activation and dropout 0.3. Models are trained from scratch with a batch size of 32 using Adam at a learning rate of 10^{-3} without weight decay. We train M1, M2, ELD, and BMSAQ for 30 epochs and EC and METR-LA for 8 epochs. Checkpoints are selected using source-validation ROC AUC only.

A.4 Hyperparameters and selection

LBNS and comparator coefficients were selected by a small manual search using source-validation performance only; target-domain observations and metrics were never used. Dataset-specific neighbourhood parameters reflect differences in sample count, sequence length, domain count, and within-class density. Table 3 reports the effective coefficients λ_{inv} and λ_{struct} used in the total objective. The CDANN coefficient was selected from $\{0.2, 0.8, 2.0\}$ and the MatchDG coefficient from $\{0.01, 0.1, 1.0\}$ using source-validation ROC AUC. DIVERSIFY uses $(K, \alpha, \alpha_1) = (2, 1, 1)$ and V-REx uses $\beta = 10$, across datasets.

Parameter	M1	M2	EC	ELD	BMSAQ	METR-LA
k	12	8	6	12	8	15
σ	0.5	1.0	1.0	1.5	0.8	0.8
λ_{inv}	0.5	0.4	1.5	0.0034	0.3	0.005
λ_{struct}	0.0025	0.04	0.6	0.0001	0.03	0.0005

Table 3: Dataset-specific LBNS hyperparameters.

For completeness, Table 4 gives the exact PR AUC values reported in Figure 2.

For paired observations, Cohen's d_z is defined as

$$d_z = \frac{\bar{\Delta}}{s_\Delta},$$

where $\bar{\Delta}$ and s_Δ are the mean and standard deviation of the paired differences, respectively. We use $|d_z| = 0.2$, 0.5, and 0.8 as conventional reference values for small, medium, and large effects.

The domain-recoverability probe freezes the selected encoder and extracts source-training and held-out source-validation representations. A separate $H \to 64 \to 32 \to D$ MLP with ReLU hidden layers is trained only on source-training representations for 15 epochs using Adam at 10^{-3} and minibatches of 128, then evaluated on source-validation representations. This source-validation accuracy is the only held-out probe score used or reported; target-domain probe accuracy is not part

Method	M1	M2	EC	ELD	BMSAQ	METR-LA
ERM	27.1±2.5	57.3±4.1	64.9±7.6	82.3±2.1	90.3±1.0	88.0±0.7
DANN	26.6±3.0	60.2±4.1	67.0±7.1	69.9±4.7	90.2±0.5	87.0±0.7
CCA	33.7±3.7	62.5±4.3	64.3±7.7	81.8±2.5	90.6±1.0	88.2±0.7
CDANN	26.6±3.1	59.1±4.5	63.1±7.8	76.0±4.0	90.5±1.0	88.0±0.8
MatchDG	24.6±1.4	62.5±5.0	65.6±6.8	82.4±2.1	**91.2±1.1**	88.1±0.8
DIVERSIFY	30.8±2.9	59.9±4.4	67.0±6.7	80.9±2.4	89.7±1.1	86.6±0.7
V-REx	28.5±3.4	64.0±4.1	69.0±6.6	**82.5±2.1**	89.9±1.1	85.7±0.9
LBNS	**40.6±4.4**	**66.3±4.8**	**78.5±4.8**	82.2±2.5	90.7±1.0	**88.5±0.7**

Table 4: Target-domain PR AUC (%) with 95% confidence intervals over 30 folds.

of the analysis. Table 2 normalises this accuracy by $1/|\mathcal{D}_{\text{source}}|$. The probe never updates the encoder, and we interpret it only as a measure of global source-domain recoverability.

A.5 Batch-size and neighbourhood-size sensitivity

We retrain LBNS on the same 30 EC folds while changing one graph parameter at a time. With $k = 6$ fixed, performance is stable from $B = 8$ to $B = 32$; the paired intervals for $B = 8$ and $B = 32$ relative to $B = 16$ configuration include zero. At least 99.1% of anchors have a selected cross-domain neighbour throughout. The larger $B = 64$ setting is modestly worse, showing that a larger candidate pool is not automatically beneficial. Varying k from 3 to 10 gives overlapping ROC intervals; $k = 3$ has lower PR than $k = 6$, consistent with a noisier local reference when the neighbourhood is undersized.

Setting	ROC AUC	PR AUC	Mean selected k	Cross-domain coverage
$B = 8,\ k = 6$	0.816	0.790	4.64	99.1%
$B = 16,\ k = 6$	0.809	0.775	5.42	99.7%
$B = 32,\ k = 6$	0.811	0.785	5.88	100.0%
$B = 64,\ k = 6$	0.792	0.734	5.99	99.9%
$B = 16,\ k = 3$	0.800	0.757	—	—
$B = 16,\ k = 10$	0.819	0.794	—	—

Table 5: EC sensitivity over the same 30 paired folds.

A.6 Representation-dimension sensitivity

To test whether LBNS depends on the fixed $H = 32$ representation, we retrain it at $H \in \{16, 64\}$ over all 30 folds of every dataset. Table 6 reports descriptive changes relative to the main-study $H = 32$. Predictive performance and domain recoverability remain broadly stable and were not significant after Holm correction (paired Wilcoxon, adjusted $p \geq 0.188$).

A.7 Conditioning-granularity controls

We compare full LBNS with two matched controls over the first ten folds of M1 and M2 at $B = 32$. *Local-uniform* retains the same hard k-NN graph and losses but removes RBF distance weighting. *Class-uniform* uses all same-label, non-self samples, changing the reference from local to class-wide. We hold the architecture, folds, initialisation, coefficients, training schedule, and source-validation checkpoint selection fixed.

Direct representation-alignment control. To isolate the effect of residualisation from local conditioning, we replace the LBNS residual-matching term with direct alignment of the complete representations while keeping the same local same-label neighbourhoods and cross-domain compari-

H	Metric	M1	M2	EC	ELD	BMSAQ	METR-LA	Average
16	ROC AUC	-2.11	-3.79	-3.01	$+0.74$	-1.41	-1.23	-1.80
	PR AUC	-1.79	-1.59	-5.17	$+0.95$	-1.02	-1.46	-1.68
	Probe accuracy	$+0.80$	$+1.23$	$+1.91$	$+4.12$	$+0.06$	$+1.92$	$+1.67$
64	ROC AUC	-2.39	$+0.77$	-0.62	-1.27	-0.15	-0.26	-0.65
	PR AUC	-3.32	$+0.63$	-2.00	-1.22	-0.12	-0.25	-1.05
	Probe accuracy	-1.87	$+0.88$	-1.02	-0.36	-0.56	$+0.14$	-0.46

Table 6: Representation-dimension sensitivity, reported as percentage-point change relative to the main-study $H = 32$ result. Positive ROC and PR changes indicate better prediction; negative probe changes indicate lower source-domain recoverability.

Dataset	Variant	ROC AUC	PR AUC
M1	LBNS (local k-NN + RBF)	0.590	0.417
	Local k-NN + uniform	0.530	0.392
	Class-wide + uniform	0.532	0.386
M2	LBNS (local k-NN + RBF)	0.673	0.687
	Local k-NN + uniform	0.642	0.676
	Class-wide + uniform	0.645	0.668

Table 7: Matched ten-fold component controls. LBNS is directionally strongest on both tasks, but the fold-bootstrap intervals for the paired differences include zero; these controls support, rather than prove, that every component is uniquely necessary.

son set $\mathcal{P}$. The resulting objective is

$$\mathcal{L}_{\text{direct}} = \mathcal{L}_{\text{task}} + \lambda_{\text{direct}} \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \|z_i - z_j\|_2^2 + \lambda_{\text{struct}} \mathcal{L}_{\text{struct}}.$$

This control therefore differs from LBNS only in the quantity being aligned. Direct alignment penalises $\|z_i - z_j\|_2^2$, whereas LBNS penalises the residual discrepancy $\|r_i - r_j\|_2^2$ with $r_i = z_i - m_i$. The neighbourhood construction, cross-domain pair set $\mathcal{P}$, encoder architecture, folds, initialisation, training schedule, source-validation checkpoint selection, and all remaining experimental settings are held fixed. For each dataset, we set λ_{direct} equal to the corresponding LBNS invariance coefficient λ_{inv}, so that the comparison changes only the aligned representation quantity.

Dataset	ROC AUC (%)	PR AUC (%)	Probe ratio
M1	53.30	33.92	1.456
M2	62.31	64.34	1.524
EC	81.21	77.80	1.389
ELD	79.80	79.91	5.512
BMSAQ	84.36	90.51	1.031
METR-LA	88.92	88.19	1.498
Average	74.98	72.45	2.068

Table 8: Results for Local Representation Alignment, which directly aligns complete representations using the same local conditioning, coefficient, and experimental protocol as LBNS. AUC values are percentages. Probe accuracy is normalised by random-chance accuracy; lower is better.

B TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

This appendix provides formal details for the theoretical motivation in Section 4. The results below are not intended to establish a general target-domain risk guarantee. Rather, they formalise two claims

Dataset	ROC AUC (%)	PR AUC (%)	Probe ratio
M1	57.88	40.08	1.755
M2	63.51	64.53	1.243
EC	79.46	73.45	1.444
ELD	80.68	80.68	5.512
BMSAQ	84.53	90.91	1.016
METR-LA	89.00	88.49	1.757
Average	75.84	73.02	2.121

Table 9: Performance of the LBNS ablation without the structure term, evaluated over the same 30 paired folds as full LBNS. All other components and training settings are unchanged. AUC values are percentages. Probe accuracy is normalised by random-chance accuracy; lower is better.

used in the main text: i) that conditioning nuisance suppression only on the observed label can be too coarse when labels aggregate multiple latent behavioural states; ii) that LBNS can be viewed as a local empirical surrogate for latent-state-conditioned nuisance suppression under a local coherence assumption.

B.1 Class-conditional invariance can be too coarse

Proposition (**Formal version of Proposition 1**). There exist distributions satisfying Assumptions 1–3 for which class-conditional invariance, $z \perp d \mid y$, is incompatible with preserving the latent behavioural state, whereas latent-state-conditioned invariance, $z \perp d \mid s$, preserves that state.

Proof. Let the latent behavioural state be $s = (y, u)$, where $y \in \{0, 1\}$ is the observed coarse label and $u \in \{0, 1\}$ distinguishes two task-relevant latent states within each label. Thus, $y = h(s)$ is many-to-one.

Consider the class $y = 0$ and two domains $d \in \{0, 1\}$. Suppose

$$\Pr(u = 1 \mid d = 0, y = 0) = p_0, \qquad \Pr(u = 1 \mid d = 1, y = 0) = p_1,$$

where

$$0 < p_0 \neq p_1 < 1.$$

For example, take $p_0 = 0.25$ and $p_1 = 0.75$. Hence, both latent states occur with positive probability in both domains, but their proportions differ.

Let the observed input be $x = (s, v)$, where v is a domain-dependent nuisance component that can be removed after conditioning on s, satisfying Assumption 3. Suppose that a representation z preserves the latent behavioural state, meaning that s is recoverable from z. If $z \perp d \mid y$, then every deterministic function of z is also conditionally independent of d given y. Recoverability of s, and hence of u, would therefore imply $u \perp d \mid y$. This contradicts $p_0 \neq p_1$. Consequently, no representation can both preserve s and satisfy $z \perp d \mid y$ in this construction.

By contrast, consider the representation $z = s$. Conditional on s, the representation is deterministic:

$$p(z \mid s, d) = p(z \mid s).$$

Thus,

$$z \perp d \mid s.$$

Because both domains have positive support for both values of u, this conditional-independence statement is non-vacuous. The representation preserves the latent behavioural state even though the state proportions differ across domains. Hence, class-conditional invariance can be too coarse under within-class state-proportion shift, whereas latent-state-conditioned invariance need not remove that structure. □

B.2 LBNS as a local surrogate under neighbourhood coherence

Proposition (**Formal version of Proposition 2**). Assume that representations decompose locally as

$$z_i = a(s_i) + n(s_i, d_i) + \epsilon_i,$$

where $a(s_i)$ is shared behavioural structure, $n(s_i, d_i)$ is a domain-dependent nuisance component, and ϵ_i is noise. Let ρ be a metric on the latent-state space and suppose that, for every cross-domain local pair $(i, j) \in \mathcal{P}$,

$$\rho(s_i, s_j) \leq \delta.$$

Assume that a is L_a-Lipschitz and that $n(\cdot, d)$ is L_n-Lipschitz on the relevant local support for every domain d. Further suppose that, for $\ell \in \{i, j\}$,

$$m_\ell = a(s_\ell) + e_\ell, \qquad \|e_\ell\|_2 \leq \eta, \qquad \|\epsilon_\ell\|_2 \leq \xi.$$

Then

$$\|(r_i - r_j) - [n(s_i, d_i) - n(s_i, d_j)]\|_2 \leq L_n\delta + 2\eta + 2\xi.$$

Thus, residual mismatch approximates within-state cross-domain nuisance mismatch up to bounded local-state, reference-estimation, and noise errors. By contrast, direct representation mismatch contains the behavioural displacement $a(s_i) - a(s_j)$, which need not vanish when $s_i \neq s_j$.

Proof. Because $m_i = a(s_i) + e_i$, the residual is

$$r_i = z_i - m_i = n(s_i, d_i) + \epsilon_i - e_i.$$

Consequently,

$$r_i - r_j = n(s_i, d_i) - n(s_j, d_j) + (\epsilon_i - \epsilon_j) - (e_i - e_j).$$

Subtracting the within-state nuisance discrepancy at s_i gives

$$\begin{aligned}(r_i - r_j) - [n(s_i, d_i) - n(s_i, d_j)] \\ = n(s_i, d_j) - n(s_j, d_j) + (\epsilon_i - \epsilon_j) - (e_i - e_j).\end{aligned}$$

By the triangle inequality, local Lipschitz continuity of $n(\cdot, d_j)$, and the bounded-error assumptions,

$$\|(r_i - r_j) - [n(s_i, d_i) - n(s_i, d_j)]\|_2 \leq L_n\delta + 2\xi + 2\eta.$$

The reverse triangle inequality therefore also gives

$$\left|\|r_i - r_j\|_2 - \|n(s_i, d_i) - n(s_i, d_j)\|_2\right| \leq L_n\delta + 2\xi + 2\eta.$$

Hence, the discrepancy penalised by $\mathcal{L}_{\text{inv}}$ approximates within-state cross-domain nuisance discrepancy up to the stated local approximation terms.

Direct representation mismatch instead satisfies

$$z_i - z_j = [a(s_i) - a(s_j)] + [n(s_i, d_i) - n(s_j, d_j)] + [\epsilon_i - \epsilon_j].$$

Its behavioural term obeys

$$\|a(s_i) - a(s_j)\|_2 \leq L_a\delta,$$

but need not be zero when the paired states are locally comparable rather than identical. Direct representation mismatch therefore does not isolate nuisance variation without additional structural assumptions. This establishes a bounded surrogate justification for the residual-suppression term; it does not claim that LBNS establishes $z \perp d \mid s$ or a target-domain risk guarantee. □